On NK-Community Problem (Theoretical Computer Science and its Applications)

AUTHOR(S):

Nakamura, Atsuyoshi; Shigezumi, Takeya; Yamamoto, Masaki

数理解析研究所講究録 1426 巻 2005 年 71-77

On NK-Community Problem

中村 篤祥 (Atsuyoshi Nakamura) * 繁住 健哉 (Takeya Shigezumi) †
山本 真基 (Masaki Yamamoto) †

January 31, 2005

Abstract

The web community is one of the structures which the World Wide Web (WWW) network has. And a network such as the WWW is represented as a graph. In this paper, we consider the following structure called *NK-community* (standing for *Nakamura-Kudo Community*) on a given simple (undirected and unweighted) graph. The NK-community is a set of vertices that link to more vertices inside the community than to vertices outside the community. We proved intractability of the *st*NK-community problem which is a variant of the NK-community problem.

1 Introduction

The rapid growth of the World Wide Web (WWW) has made more information freely available than ever before. In the WWW, web sites are linked to each other so that they are referred to their related sites. It would be more useful if *communities*, that is groups of individuals which share a common interest, were identified.

Consider, for example, search engine crawlers sample the indexable web often enough to insure that results are valid, and broadly enough to insure that all valuable documents are indexed. However, it doesn't seem to be practical accroding to the fact [4] that no search engine covers more than about 16%, and the union of eleven major searh engines covers less than 50%.

For another example, in information retrieval, there's a classic tension between recall and precision. Specifying more recall (trying to find all the relevant items), you often get a lot of junk. If you limit your search trying to find only precisely relevant items, you can miss important items because they don't use quite the same vocabulary.

By those reasons, a notion of *web community* was introduced in [1]. It may enable web crawlers to effectively focus on narrow but topically related subsets of the web, and

*北海道大学大学院情報科学研究科 Graduate School of Information Science and Technology, Hokkaido University
†東京工業大学大学院情報理工学研究科 Dept. of Mathematical and Computing Sciences Tokyo Institute of Technology

also enable search engines to increase the precision and recall of search results. They define a community to be a set of web sites that are linked to more web sites inside the community than to web sites outside the community. Specifically, regarding web sites as vertices V and links as (undirected) edges E, a subset C of V is a community iff,

$$\forall u \in C \left[\left| \bigcup_{v \in C} \{u, v\} \right| \geq \left| \bigcup_{v \in V \backslash C} \{u, v\} \right| \right] \tag{1}$$

On the other hand, Nakamura and Kudo gave different definitions of communities, pointing out the uncertainty of their definition (see [2]): weak NK-community, NK-community, and *st*NK-community. Given an undirected graph $G(V, E)$, a community $C \subset V$ is an NK-community iff C satisfies a stricter condition

$$\forall u \in C \left[\left| \bigcup_{v \in C} \{u, v\} \right| > \left| \bigcup_{v \in V \backslash C} \{u, v\} \right| \right], \tag{2}$$

and $V \backslash C$ satisfies the condition (1) above. An weak NK-community requires that both C and $V \backslash C$ satisfy the condition (1). In particular vertices s and t, an *st*NK-community is an NK-community such that $s \in C$ and $t \in V \backslash C$. They state in the concluding section as a future work that to analyse the hardness of finding *st*NK-community will promote new algorithm for the NK-community problem.

In this paper, we solve this problem: we'll show that *st*NK-community is NP-complete. The rest of paper is organized as follows. Some definitions and preliminaries are described in Section 2. In Section 3, we prove this problem is NP-complete. Finally, we give some concluding remarks and future works in Section 4.

2 Preliminaries

We consider the following structure called *stNK-community* (standing for *Nakamura-Kudo Community*) on a given simple (undirected and unweighted) graph. In particular, we consider about the hardness of the following problem:

<u>*st*NK-community Problem</u>

Instance: A simple (undirected and unweighted) graph $G = (V, E)$, and a pair (s, t) of vertices of V.

Question: Is there any partition (S, T) (i.e., $S \cap T = \phi$ and $S \cup T = V$) such that $s \in S$, $t \in T$, and the following two are satisfied?

$$\forall u \in S [\ |\{\{u, v\} \in E : v \in S\}| > |\{\{u, v\} \in E : v \in T\}|\ ],$$
$$\forall u \in T [\ |\{\{u, v\} \in E : v \in T\}| \geq |\{\{u, v\} \in E : v \in S\}|\ ].$$

Moreover, we consider a variant of *st*NK-community problem as follows.

Weak stNK-community Problem

Instance: A simple (undirected and unweighted) graph $G = (V, E)$, and a pair (s, t) of vertices of V.

Question: Is there any partition (S, T) (i.e., $S \cap T = \phi$ and $S \cup T = V$) such that $s \in S$, $t \in T$, and the following two are satisfied?

$$\forall u \in S\,[\,|\{\{u,v\} \in E : v \in S\}| \geq |\{\{u,v\} \in E : v \in T\}|\,],$$
$$\forall u \in T\,[\,|\{\{u,v\} \in E : v \in T\}| \geq |\{\{u,v\} \in E : v \in S\}|\,].$$

Given a graph $G = (V, E)$, we say that a vertex set $S \subset V$ is an *(weak) stNK-community* if such a partition (S, T) as above exists.

Finally, we define a useful notation to describe stNK-community conditions.

Definition 2.1

$$\forall v \in S, \quad GAP(v) = |\{\{u,v\} : u \in S\}| - |\{\{u,v\} : u \in T\}|,$$
$$\forall v \in T, \quad GAP(v) = |\{\{u,v\} : u \in T\}| - |\{\{u,v\} : u \in S\}|.$$

3 Hardness of (Weak) stNK-community

In this section, we'll show both stNK-community problem are NP-complete as follows: We first reduce 3-SAT to a variant of 3-SAT, and then the variant of 3-SAT is reduced to stNK-community problem. Note that this reduction is suitable for both of the stNK-community problem and weak stNK-community problem. So we'll simply call that community stNK-community when there is no confusion.

NOIT (No-One-In-Three) 3-SAT Problem

Instance: A 3-CNF formula F over X.

Question: Is there any assignment σ to X such that no clause of F has exactly one true literal under σ?

Lemma 3.1 NOIT 3-SAT is NP-complete. ■

Theorem 3.2 (Weak) stNK-community problem is NP-complete.

Proof. Given a 3-CNF formula F over $X = \{x_1, \cdots, x_n\}$ for NOIT 3-SAT, we construct a graph $G = (V, E)$ for stNK-community problem as follows: Indeed, G is composed of three parts of graphs $G_1 = (V_1, E_1)$, $G_2 = (V_2, E_2)$, and $G_3 = (V_3, E_3)$, that is, $V = V_1 \cup V_2 \cup V_3$ and $E = E_1 \cup E_2 \cup E_3$.

Construction: Let $occ(l)$ be the number of occurrences of literal l, and let $L = \{l : l \in X \text{ or } \bar{l} \in X\}$ and

$$k \stackrel{\text{def}}{=} 2 + \max\{n, \max\{occ(l) : l \in L\}\}.$$

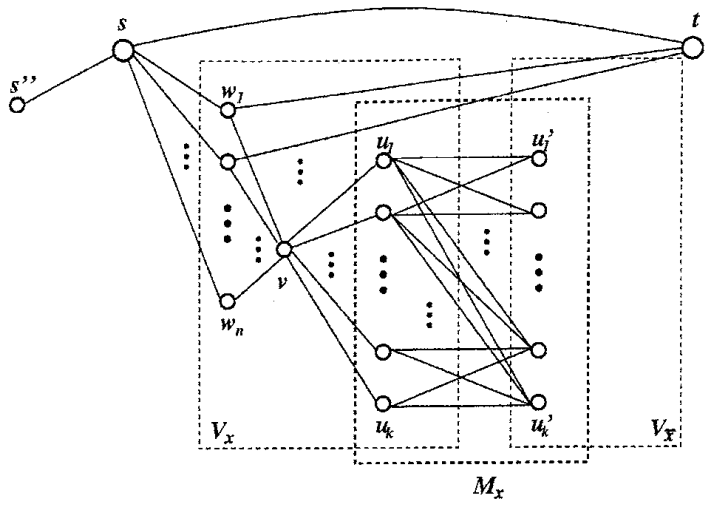

Figure 1: $G_1 = (V_1, E_1)$

We explain the construction of G_1 presented in Figure 1. For each variable x of X, we first construct two parts made up from vertex sets $V_x \cup \{s,t\}$ and $V_{\bar{x}} \cup \{s,t\}$, respectively, which are identical and associated with x and $\bar{x}$, respectively. For the part of $V_x \cup \{s,t\}$, k vertices $u_1, \cdots, u_k$ and n vertices $w_1, \cdots, w_n$ are all connected to a vertex v. Moreover, each of $w_1, \cdots, w_n$ is connected to both of s and t. We finally make a clique M_x consisting of vertices $u_1, \cdots, u_k$ and $u'_1, \cdots, u'_k$. (In the figure, only edges between $u_1, \cdots, u_k$ and $u'_1, \cdots, u'_k$ are shown.)

Next, we explain the construction of G_2 presented in Figure 2. In this figure, all the clique parts $M_{x_1}, \cdots, M_{x_n}$ which appear in G_1 are shown again. Every clique part, say M_{x_1} associated with a variable $x_1 \in X$, is connected to a vertex s' through the last two vertices x_{1k} and x'_{1k} as shown in the figure. We call vertices $y_1, y'_1, \cdots, y_n, y'_n$ in the figure, *extra* vertices for $\{x_{1k-1}, x_{1k}\}, \{x'_{1k-1}, x'_{1k}\}, \cdots, \{x_{nk-1}, x_{nk}\}, \{x'_{nk-1}, x'_{nk}\}$, respectively. A clique T' is composed of $t_1, \cdots, t_n$. The vertex s' is directly connected to s, and is also connected to t through $t_1, \cdots, t_n$.

Finally, we explain the construction of G_3 presented in Figure 3. In this figure, only one part of G_3 which corresponds to, say a clause $C = (x_1 \vee \overline{x_2} \vee x_3)$, is shown as an example. We first construct a clique composed of c, a_1, a_2, a_3, which has four dummy vertices as shown in the figure, and then we connect each of vertices a_1, a_2 and a_3 to x_{11}, x'_{21} and x_{31}, respectively. We call the vertex c as clause vertex. Finally, we construct the vertex s'' as shown in the figure. Each of $c_1, \cdots, c_m$ is connected to s'', and s'' is connected to s. Moreover, s'' has m dummy vertices. If literal x_1 appears i times in F, each of $x_{11}, \cdots, x_{1i}$ of M_{x_1} is connected to some triangle which x_1 appears in. Observe that we never have $i \geq k$. The rest except for x_{1k} (i.e., $x_{1i+1}, \cdots, x_{1k-1}$) are connected to dummy vertices for each. (We do the same construction for the other literals of L.) We call the vertex, say a_1, of the clique connected to u_{11}, *extra* for u_{11}, and also each of those dummy vertices connected to $u_{1i+1}, \cdots, x_{1k-1}$, *extra* for $u_{1i+1}, \cdots, x_{1k-1}$,

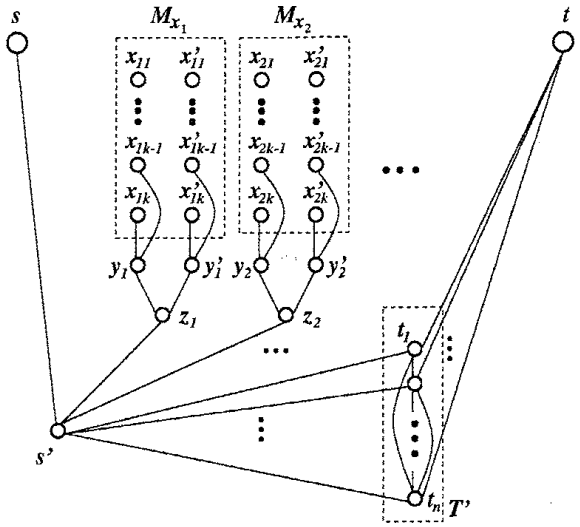

Figure 2: $G_2 = (V_2, E_2)$

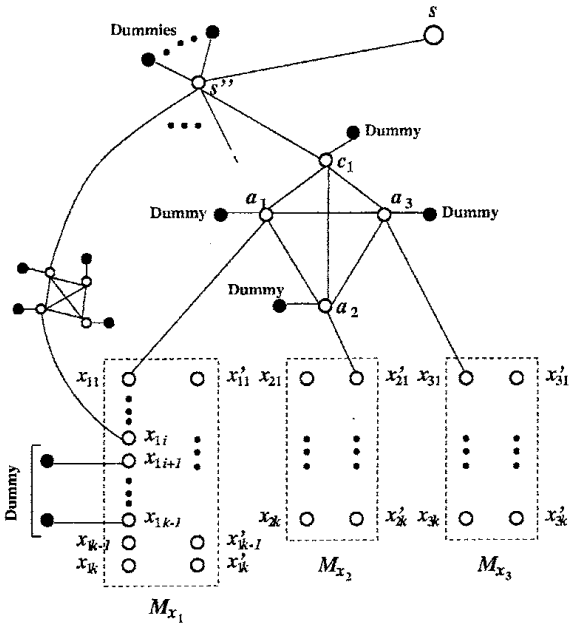

Figure 3: $G_3 = (V_3, E_3)$

respectively.

Consistency: We prove that F is satisfiable in the sense of No-One-In-Three if and only if G constructed above has an stNK-community S. Let $T = V \backslash S$. We first observe a few necessary conditions for the existence of stNK-community , which are independent of the given formula F. Observe first that for each $x \in X$, the set of vertices $w_1, \cdots, w_n$, and v cannot be partitioned because for each i, it is impossible that vertices w_i and v are divided. We further have the following claim about vertices $u_1, \cdots, u_k$:

Claim 1 It is necessary that for every clique M_x, the set U of k vertices $u_1, \cdots, u_k$ is not partitioned, and neither is the set U' of k vertices $u'_1, \cdots, u'_k$.

Proof of the claim: Suppose that there is an stNK-community S such that at least one of U and U' is partitioned into S and T. We assume w.l.o.g. that U is partitioned so that i $(0 < i < k)$ vertices $u_1, \cdots, u_i$ are in S, and $(k-i)$ vertices $u_{i+1}, \cdots, u_k$ are in T. Similarly, j $(0 \leq j \leq k)$ vertices $u'_1, \cdots, u'_j$ are in S and $(k-j)$ vertices $u'_{j+1}, \cdots, u'_k$ are in T. Suppose further that $v \in S$. (It is similarly proved for the case of $v \in T$.) Thus, there exist two vertices, say $u_1 \in S$ and $u_k \in T$, such that $\mathrm{GAP}(u_1) > 0$ and $\mathrm{GAP}(u_k) \geq 0$. It is easy to see that $\mathrm{GAP}(u_1) = (i-1)+j+2-(k-i+k-j) = 2(i+j)-2k+1 > 0$, for the case that the extra vertex for u_1 is in S, and $\mathrm{GAP}(u_k) = (k-i-1)+(k-j)+1-(i+j+1) = 2k-2(i+j)-1 \geq 0$, for the case that the extra vertex for u_k is in T. (Note that it suffices to show for these cases.) These two inequalities above lead to a contradiction because we have $i+j \geq k$ from the first, and $k \geq i+j+1$ from the second.

Summing up the above mentioned, we have that for each $x \in X$, the set W of $w_1, \cdots, w_n$, and v cannot be partitioned, and neither can the set U of $u_1, \cdots, u_k$. It follows that the set $U \cup W$ cannot be partitioned because $n < k$ and therefore v must be in the same partition as U. Thus, for every $x \in X$, all vertices of V_x must be in the same partition, so must those of $V_{\bar{x}}$. We now claim the following.

Claim 2 *All vertices of clique T' are in T.*

Proof of the claim: First, we prove that all vertices of clique T' are always in the same partition. If T' is partitioned into different partition, e.g. $t_1, \cdots, t_l \in S$ and $t_{l+1}, \cdots, t_n \in T$, we have

$$\begin{aligned} \mathrm{GAP}(t_1) &= 1 + l - 1 - (n - l + 1) < 0 \quad \left(\text{if } l \leq \left\lfloor \frac{n}{2} \right\rfloor\right) \\ \mathrm{GAP}(t_n) &= 1 + (n - l - 1) - (l + 1) < 0 \quad \left(\text{if } l > \left\lfloor \frac{n}{2} \right\rfloor\right). \end{aligned}$$

This contradicts that T' is partitioned into different partition. Suppose that for $X' \subset X$ and $X'' \subset X$ such that $X' \cap X'' = \phi$, for every $x \in X'$ all vertices of $V_x \cup V_{\bar{x}}$ are in S, and for every $x \in X''$ all vertices of $V_x \cup V_{\bar{x}}$ are in T. Let $|X'| = i$ and $|X''| = j$. And we assume all vertices $t_1, \cdots, t_n$ are in S, we have $\mathrm{GAP}(t) = 2nj - (1 + 2ni + n) = 2n(j - i) - 1 - n \geq 0$. From this we obtain $j > i$. Since $t_1, \cdots, t_n$ are in S, s' must be in S. Thus, we have $\mathrm{GAP}(s) = 2ni - (2nj + 1) = 2n(i - j) + 1 > 0$. From this we obtain $j \leq i$, this is a contradiction.

We now claim the following which implies the correspondence of a partition of V to an assignment to X.

Claim 3 For every $x \in X$, all vertices of V_x are in S iff all vertices of $V_{\bar{x}}$ are in T.

Proof of the claim: Suppose that X', X'', i and j are same as the proof of claim 2. Because of claim 2, vertices $t_1, \cdots, t_n$ must be all in T, we have $\mathrm{GAP}(t) = n + 2nj - 2ni = n + 2n(j - i) \geq 0$. From this we obtain $j \geq i$. We now assume $j > 0$, e.g., $V_{x_1} \cup V_{\bar{x_1}} \subset T$. Then, s' must be in T because y_1, y_1', z_1 must be in T (see Figure 2), and therefore the number of vertices which is in T and adjacent to s' is at least $n + 1$ while the number of vertices which is in S and adjacent to s' is at most n. Thus, we have $\mathrm{GAP}(s) = 2ni - 2nj - 1 = 2n(i - j) - 1 > 0$. From this we obtain $i > j$. This contradicts to $j \geq i$, therefore we have $j = 0$. Immediately, we also have $i = 0$ because of $j \geq i$.

From this claim, s' must be in S, which follows that all vertices $z_1, \ldots, z_n$ in Figure 2 must be in S. Moreover, all the extra vertices must be in the same partition as vertices in M_x adjacent to those extra vertices. Note that those conditions on S all mentioned above must be satisfied whatever the given formula F is.

We now show the relationship: F is satisfiable in the sense of No-One-In-Three if and only if the *st*NK-community condition on S is satisfied at all the vertices of triangles in

G_3. Recall that, for example, x_{11} and the extra vertex a_1 in the triangle must be in the same partition. Thus, we can regard the partition (S, T) as an assignment to X: that is, the variables x such that V_x is in S is assigned true, and the variables x such that V_x is in T is assigned false. Consider an arbitrary assignment σ to X. According to the assignment σ, we have four types of the placement of triangles: 1) all the three vertices of a triangle are in the same partition, 2) two vertices out of the three are in S, and 3) two vertices out of the three are in T. For the cases 1) and 2), the *st*NK-community condition is satisfied at all the three vertices, which corresponds to the clauses satisfied by σ. On the other hand, for the case 3), the *st*NK-community condition is not satisfied at a vertex in S, which corresponds to the clauses not satisfied by σ. ■

4 Summary and future work

The *st*NK-community problem is introduced by Nakamura and Kudo [2]. We've proved this problem is NP-complete.

Since the *st*NK-community problem is NP-complete, we consider about the randomized algorithm to solve it with high probability. And we consider about the suitable definition of the optimization problem and the approximation algorithm for it. Perhaps, we need to make a simpler reduction in the proof of the NP-hardness of optimized version of *st*NK-community problem to consider the approximation algorithm. The purpose of this problem is mining the communities of web structure. We also consider about the NK-community problem whose instance doesn't fix (s, t). Flake, Tarjan and Tsioutsiouliklis [3] proved NP-completeness for this problem in case of G has edge weight function $w : E \to \mathbf{Z}^+$ and the community condition inequality is $\forall u \in S, \left[\sum_{\{\{u,v\} \in E : v \in S\}} w(\{u,v\}) > \sum_{\{\{u,v\} \in E : v \in T\}} w(\{u,v\}) \right]$. But, it's not trivial to translate their proof to the case of the weights restricted to be 1.

References

[1] Gary William Flake, Steve Lawrence, C. Lee Giles. Efficient Identification of Web Communities. *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2000), 150-160.

[2] 中村篤祥, 工藤峰一. コミュニティトポロジーによる Web グラフ分割. 第4回データマイニング ワークショップ, (2004).

[3] Gary William Flake, Robert E. Tarjan, Kostas Tsioutsiouliklis. Graph Clustering and Minimum Cut Trees. *Internet Mathematics*, Vol. 1, No. 4 (2003), 385-408.

[4] Steven Lawrence and C.Lee Gliles, "Accessibility of information on the web", *Nature*, **400(6740)**, 1999, 107-109.